Game Sale Report

STA302 - Final Assignment

YUANCHEN LING

December 10, 2022

Introduction

Nowadays, Gaming has no doubt became one of the most popular hobbies worldwide. Nintendo, Electronic Arts, Rockstar games, Ubisoft entertainments are examples of big companies growing fast in the late 1990s to now because of game they produce. However, with game industry becoming more popular, the competitions in the market is also getting more intense as well. Therefore, it is reasonable for companies to predict in which way selling the game could help the company get the most benefits. The goal of this research is to find which factors would influence the sales to a large extent by using linear models. After doing so, companies could predict or improve their marketing strategy for bigger benefits.

It is reasonable for people to think quality is the fist key element effecting game sales. Logically, I could consider game as a regular product in the market. With a research conducted, it finds out that the star rating(media rating or user rating) has a positive effete on game sales, which means higher the star rating is, higher the sale would be. It could be said that rating has a correlation with the sale of a product. Therefore, in this scenario, the professional game media's comments on a game might be a effect the game sale positively.[1]

And also, types of platforms might also be a key in effecting the game sales. A research is conducted by choosing the data set of game sales between 2006 to 2011 shows that clear differences exist in across different platforms.Nintendo was the biggest leader in gaming, with the Wii being the main console and the DS as the main handheld platform. Nintendo had fallen behind SONY before the sample was studied, but the pioneer of body movement and mobility in the Wii human-machine interface pushed Nintendo to the top of the list. The game sales might also relate to the format it releases to. Therefore, types of format might still be a factor of influencing the game sale till 2016.[2]

Last but not least, in my research, publication and production company might also be key factors effecting the sale of a game. The research found that small companies are more vulnerable and lack of strength in financing, which means they have less resources to propaganda their products or they lack the technology to produce the game, which surely would be a disadvantage in competing with bigger companies I have mentioned above. Thus, scale of companies is also a part I concern.[3]

To conclude, in this study, I will conduct a prediction study by performing linear model to help me find out which factors would effect the game sales significantly.As I discussed above, I may predict that the production company, platform and media score might be three important factors influencing the game sales. Therefore. my null hypothesis is that the company, platform and media score are significant in effecting the game sales. My alternative hypothesis is that the company, platform and media score are not significant in effecting the game sales. One thing to mention is that the global game sales would be calculated as a sum of all sales in different regions. It is no doubt that regions would pose a significant effect on the game sales. Accordingly, I build the model also to see which region would contribute the most to the global sales.

Method

The goal of my research is to find the best fitted linear model to predict the global sales with some important variables. To fit the model, I firstly drop the missing information, which gives me a cleaned data set.To start, I split the cleaned data set into two parts, which is training data set and test data set containing 75% and 25% data respectively. After I build these two models, I use figures for categorical and numerical variables in both data sets for EDA(Exploratory Data Analysis) and to have a general idea about the information about different kinds of variables.

Then, I build the full model, which represents the linear relationship between global sales and all the related variables in the data set. Then, since some of the numerical variables in my data set contain 0, I conduct transformation by adding 0.1 on numerical variables of training and test data sets to build the future model. Besides, I use BOXCOX transformation on the numerical variables to deal with the potential violations 1 or 2 in both full models. I would round the power accordingly to its output. That is saying if the number is 1.6, I would round it to 2 on my own. The reason for not checking residual plots for violation of assumptions is that I would do them after doing transformation in the full model. After doing the transformation, I get two similar transferred models on the training set and test data set.

Then, I use VIF to check if there are any variables with multicollinearity is not present in the model. If there exists variables with VIF>5, which means these predictors would have multicollinaerity with other variables. Accordingly, I will remove the predictors with VIF < 5 one by one. As a result, I build up an new model by adding or removing such variables manually. The remaining variables are without collinearity. After that, I reduce each variable one by one manually to check which final model fits best among all models. To distinguish which models fits better, I use ANOVA to test. If the p-value is less than 0.05, it shows that my transferred full model has a better fit compared to my full model. Removing the variable one by one and compare each model with my final model, I repeat this process and find my original full model has the best fit linearity among all models.Therefore, I may choose final model as my model.

However, this method is under all assumptions being satisfied. If assumptions are violated, my method would be less reliable. If condition1, condition2 and other assumptions are satisfied, I could ensure that my final model is the best among all models. I check additional two conditionals and residual plots to check for assumptions. I conduct this process on both test and training data set to include all the possible violations.

To check additional two conditions are violated or not, I use residual plot to check condition 1 is violated or not. By using the residual plot between predicted value for global sales(y) and predicated global sales($\hat{y}$), I may say condition 1 is not satisfied if no linear pattern exists in the figures. If there exists linear or random pattern is found between all numeric variables, I may say condition 2 has been satisfied. Next, I check other assumptions by using residuals plots and QQ plots. The residual between fitted model and residual is to check linearity assumption. Similar as above, random patterns or linear patterns represent linearity assumption being satisfied. Other residuals plots between residual and numerical variables are to check the independent assumption and homoscedasticity. If the plot shows no clusters in plots, then the independent assumption would be satisfied. If points on the graph show random patterns, then the homoscedasticity assumption would be satisfied. QQ plot is to check whether assumption normality of Errors is satisfied or not. I use residuals plots between residual and fitted models to see if the linearity assumptions is satisfied. And It is no doubt that I want the perfect model with no assumptions are being violated. However, it is not often the case in reality. Therefore, I would choose the models with less violations by building the new model manually(adding or removing variables in the previous steps).

Lastly, I check the outliers and influential points in the models. I use DEFITS as a standard to measure how many influential points would be in my model. If there are reasons for me to remove such data, I may refit an new model based on that. If not, I will take a look at both data sets and to see if there exist any new violations. If both datasets see the similar trend and linear relationship, then they would be my final models.

Results

The data set is from Kaggle.com.[4] There are 16719 games in total, containing the information about the name, publisher, developer, year of release, the media score and review counts, the user score and review counts, rating and year of release. Under my consideration, I exclude the information name of the game to fit the products since I mainly focus on the game sales. Table 4 in appendix contains the important variables which I would use for fitting the fact. Similarly, although game sales might increase with a increase in purchasing powers after year, it is hard and unrealistic for companies to control the year of release. Therefore, these two variables are not under my consideration. For the numerical variables, I multiple sales by 100 as the sale variables are in unit of billion. After I drop the missing values, I get a cleaned data set with 6826 variables. For the testing and training data set, there are 1706 and 5120 observations.

Figure1: The categorical variables on training dataset

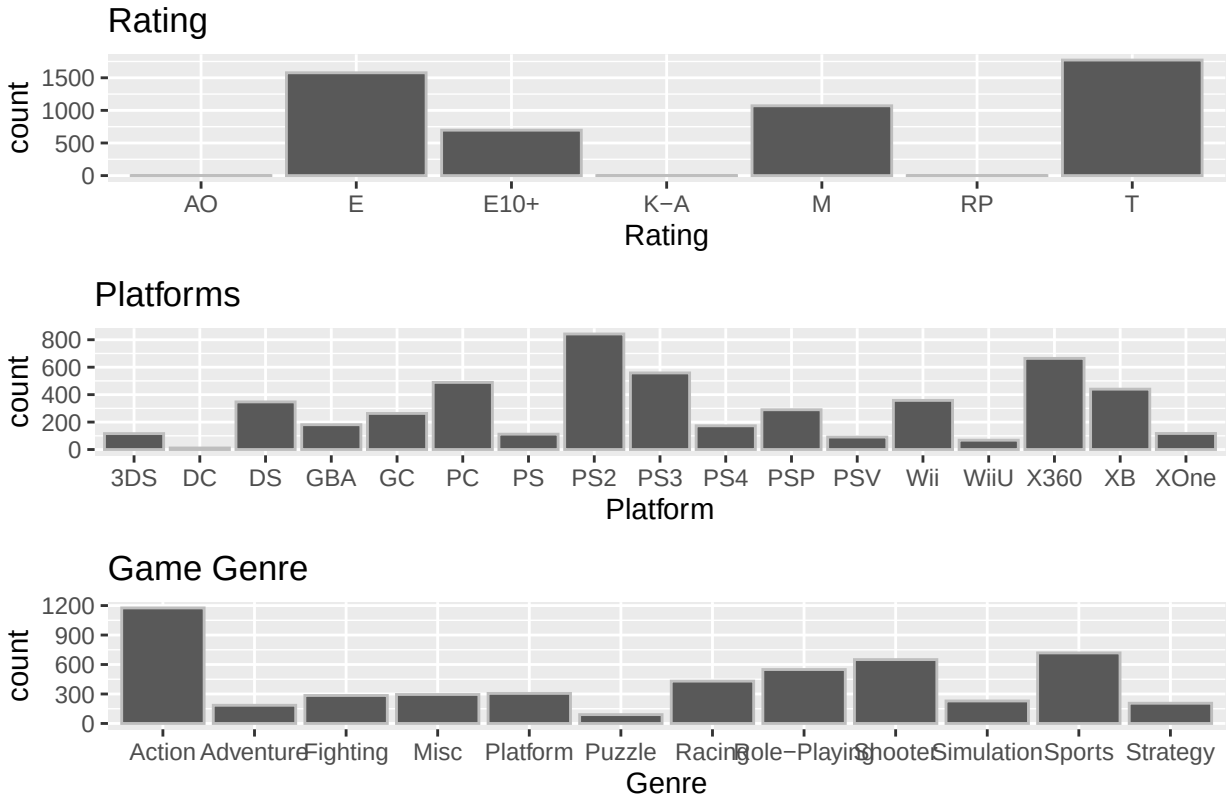


Figure2: The categorical variables on testing dataset

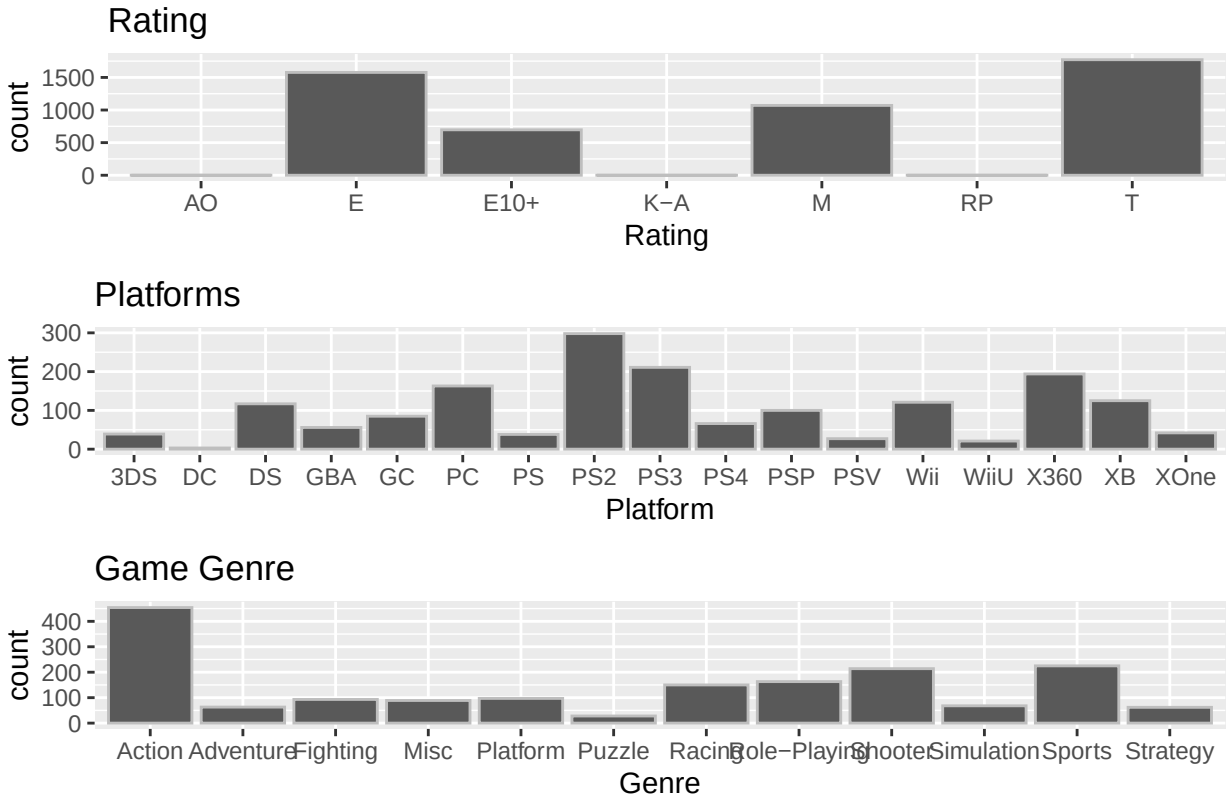


Table 1: The numerical variables of the training set

Variables	Mean	Median	Maximum	Minimum	Standard.Deviation
Sales in NA	39.530469	15.0	4136.0	0.0	103.536079
Sales in Europe	24.027930	6.0	2896.0	0.0	74.215205
Sales in Japan	6.772266	0.0	650.0	0.0	30.751150
Global Sales	78.762695	29.0	8253.0	1.0	212.781438
Media score	70.300781	73.0	98.0	19.0	13.913898
Media comment number	29.222461	25.0	113.0	113.0	19.389398
User Score on the game	7.200977	7.5	9.5	0.5	1.431433
Count of users scoring about the game	175.549805	27.0	10665.0	4.0	575.445499

Table 2: The numerical variables of the testing set

Variables	Mean	Median	Maximum	Minimum	Standard.Deviation
Sales in NA	39.182298	15.0	970.0	0.0	72.605689
Sales in Europe	22.343494	6.0	631.0	0.0	48.675975
Sales in Japan	5.342321	0.0	439.0	0.0	21.660132
Global Sales	74.710434	29.0	1461.0	1.0	135.464456
Media score	70.172333	72.0	98.0	13.0	13.742699
Media comment number	28.056858	24.0	103.0	103.0	18.692064
User Score on the game	7.137104	7.5	9.6	0.7	1.467649
Count of users scoring about the game	172.341735	25.0	10179.0	4.0	622.028587

It could be seen from both figure 1, figure 2 and table1, table2 that the distribution of different categorical and numerical variables follows a similar trend for both data sets. No significant differences exist between two data sets. For figure 1 and 2, it is pretty clear that most game are sold in PS2 platform. And most of games companies sell are action games with rates mainly being E(suitable for children older than 6) and T(T games are meant for teenagers.). It gives me a general idea about which platform is the most popular in the past decades and what kind of game is most common in the market. From the figure 3 and 4, it could be seen that the mean for global sales is around 72. Its mean is much higher than its median, which means it is super right-skewed. One interesting fact to notice is the maximum and minimum for sales in two different data sets. It is mainly due to large variance of different game sales in the original data set. Although I may say that these two groups see a significantly different trend in maximum and minimum, these two data sets could still work since the real world data usually does not follow a typical normal distribution.

By conducting the methods above, I firstly use BoxCox and remove the multi collinearity variables. After removing the variables whose VIF>5, I have a full model. In both testing and training data sets, I obtain a transferred full model, represents the linear relationship between $\log Y$(Y represents the global sales) and $log(NASales)$, $log(EuSales)$, $JPSales^{-1}$, $log(OtherSales)$, $JPSales^{2}$, $log(MediaComments)$, $(UserCount)^{3}$ and $log(UserCommentCounts)$. Then, based on above methods, I find out my global sales and predicted global sales have a strong linear pattern for both testing and training sets. It is clear that condition 1 has been satisfied in both data sets. From the fitted models in test and training data sets, I find either variables are in linear relationship or they are completely in random patterns except the variable user comments count. When checking other assumptions by residual plots, I find one variable user count has colinearity with other variables. I remove this variable and refit this model. Although full model has a better fitness, it violates the condition 2, which motivates me to fit a new model. Using it to compare other models, I find out VIF of it is still the lowest with other reduced models. After checking the model assumptions, I find out that although two conditions are satisfied, but some of other assumptions are not satisfied. By the figure5 and figure 6(See appendix), it is clear that QQ plots shows normality is violated in both data sets since the line is not a perfect linear relationship. Its variance is larger than what I expect, thus it violates the normality is not satisfied perfectly. Compared to my original preferred model, this model has a similar trend in linearity and homoscedasticity on training and testing data sets. It is likely for the model to violate homoscedasticity on some of the variables and to violate the independent assumption slightly. However, since other models still violate the assumptions, this graph has the least violation of assumptions and have the best fitness. Therefore, I still use this model for interpretation in the discussion.

Therefore, I build a model whose predictors are all less than 5. There are 476 leverage points, 41 outliers and 234 influential points in the training set. There are 548 leverage points, 14 outliers and 172 influential points in the test set. There is no strong reason for me to remove such data and no more new assumptions are violated. Therefore, I build up my final model to interpret.

Table 3: The summary about the final model

Variable	Coefficient(Train)	Coefficient(Test)	Standard Error(Train set)	Standard Error(Test)
Intercept	2.6851825	2.3055884	0.3977907	0.0613703
E rating(for everyone)	-0.2566511	0	0.3955345	0
E10+ rating(for 10+ all)	-0.3471973	-0.0582465	0.3957285	0.0329558
K-A rating(kids to adults)	0.4485159	0	0.5592928	0
M rating(for 17+)	-0.4228189	-0.1056832	0.3955659	0.0299891
RP rating(for mature 17+)	-0.3695171	0	0.5595503	0
T rating(for 13+ teenagers)	-0.3496994	-0.0771066	0.3955385	0.0258568
$logNaSales$	0.3000708	0.3280905	0.0042615	0.0075821

Variable	Coefficient(Train)	Coefficient(Test)	Standard Error(Train set)	Standard Error(Test)
$logEuSales$	0.1431731	0.1488147	0.0038189	0.0069113
$JPsales^{-1}$	-0.159668	-0.1659765	0.0048488	0.0088942
$logOtherSales$	0.2980572	0.2535331	0.0054303	0.0098533
$MediaScore^2$	5.6441296×10^{-5}	8.5119068×10^{-5}	4.0483901×10^{-6}	7.5837602×10^{-6}
$UserScore^3$	-6.167756×10^{-6}	$-1.8746117 \times 10^{-4}$	3.5735881×10^{-5}	6.5336371×10^{-5}
$log(MediaComments)$	-0.0203945	-0.0223556	0.0092648	0.0166419

Discussion

Based on Table 3, I have my full model, its linear relationship is expressed(Y represents global sales):

$$logY = \beta_0 + \beta_1 ERating + \beta_2 E10^+ Rating + \cdots + \beta_7 logNaSales + \beta_8 logEuSales + \beta_9 JPsales^{-1} + \beta_{10} logOtherSales + \beta_{11} MediaScore^2 + \beta_{12} UserScore^3 + \beta_{13} logMediaComments$$

I could see from above, test and train data set does not witness a huge difference except for test data set missing information for few variables and the variances between few variables(See Discussions). With the similar trend, I use training data to illustrate my conclusion. One thing to mention is that the test data contains less data for losing rating categories(E, K-A, RP ratings). Therefore, I mainly use training data set to illustrate the linear relationship. With the full model, it is clear that rating in this model effects the model slightly. If a game has rating K-A, it would increase for 0.4485159 in $\log GlobalSales$ if other variables are fixed. Among all regions, NA sales would account for the most part of sales. For one unit increase in $\log NaSales$, it is expected that the log of global sales would increase by 0.3000708 fixed other variables. One interesting fact is that I notice that the $\log GlobalSales$ would decreases by -0.159668 with increase in a unit increase in $JPsales^{-1}$. It is pretty shocking that the model reflects such fact. At the same time, the sales in other regions is similar significant to NA regions. The media score and media comments would also have a positive influence on the global sales. I expect the $\log GlobalSales$ would increase by 5.6441296×10^{-5} and -0.0203945 respectively, one unit increase in $MediaScore^2$ and $log(MediaComments)$. From the above results, my null hypothesis has to be rejected unfortunately. In the data set, my null hypothesis suggests that the company, platform and media score would significantly effect the global sale logically, but they are co-linear to other variables, which has been eliminated while I am trying to fit the model. It turns out the NA,Europe,Japan regions, media score and media comment numbers would effet the global sales significantly. Companies could make the corresponding market strategy for the bigger benefits.

In the test data set, some of the rating information is not included, which means some important information when splitting the data is missing. While I am trying to refit the model, the model might not be perfectly accurate.The similar trend could be found on the standard deviation for some variables(Rating) in sets. They witness a huge difference in standard deviations. Therefore, I am expecting larger data sets so that the experiment could be more accurate. And also, I drop all the missing value in the very beginning, which means some useful information might not be captured, resulting in the incorrectness of the final model. My final model does not satisfy the normality, resulting the result of the experiment less reliable.. The slight violation of dependent assumption and homoscedaisticy would also make the final model less reliable.The fact I discussed above might be resulted by BOXCOX transformation. It is not logical that Japanese sales would pose a negative influence on the global sale slightly. The power transformation might not be perfectly suitable in this scenario. There may be some extreme values in the data, which may affect the stability and accuracy of the experimental results. And in the final model, there still exists some variables with >0.05 p-value, which means they are not very significant.

Appendix

Table 4: The variables in the data set

Important_Variables	Variable_types	Variable.Disription
Platform	Cateogrical	The platform game relases
Year_of_Release	Numerical	Year of game release
Genre	Categorical	Genre of the game(action,puzzle..)
Publisher	Categorical	The publisher company of the game
NA_Sales	Numerical	Sales in North America
EU_Sales	Numerical	Sales in Europe
JP_Sales	Numerical	Sales in Japan
Other_Sales	Numerical	Sales in other regions
Global_Sales	Numerical	The sales in global
Critic_Score	Numerical	Aggregate score compiled by Metacritic staff
Critic_Count	Numerical	The numer of comments by Metacritic staff on the game
User_Score	Numerical	The score users rate on the game
User_Count	Numerical	Number of users who gave the userscore
Developer	Categorical	The developer of the game
Rating	Categorical	The ESRB ratings

Figure 5: The residual plots and normal QQ plot in training set

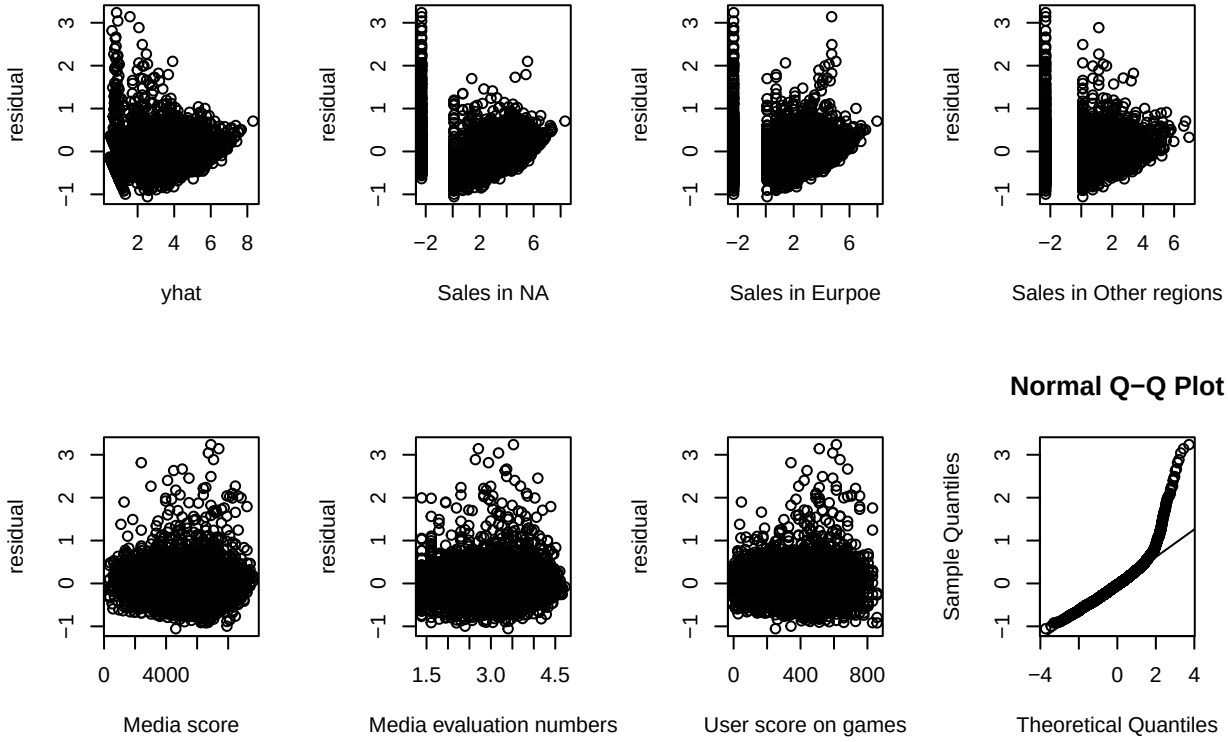
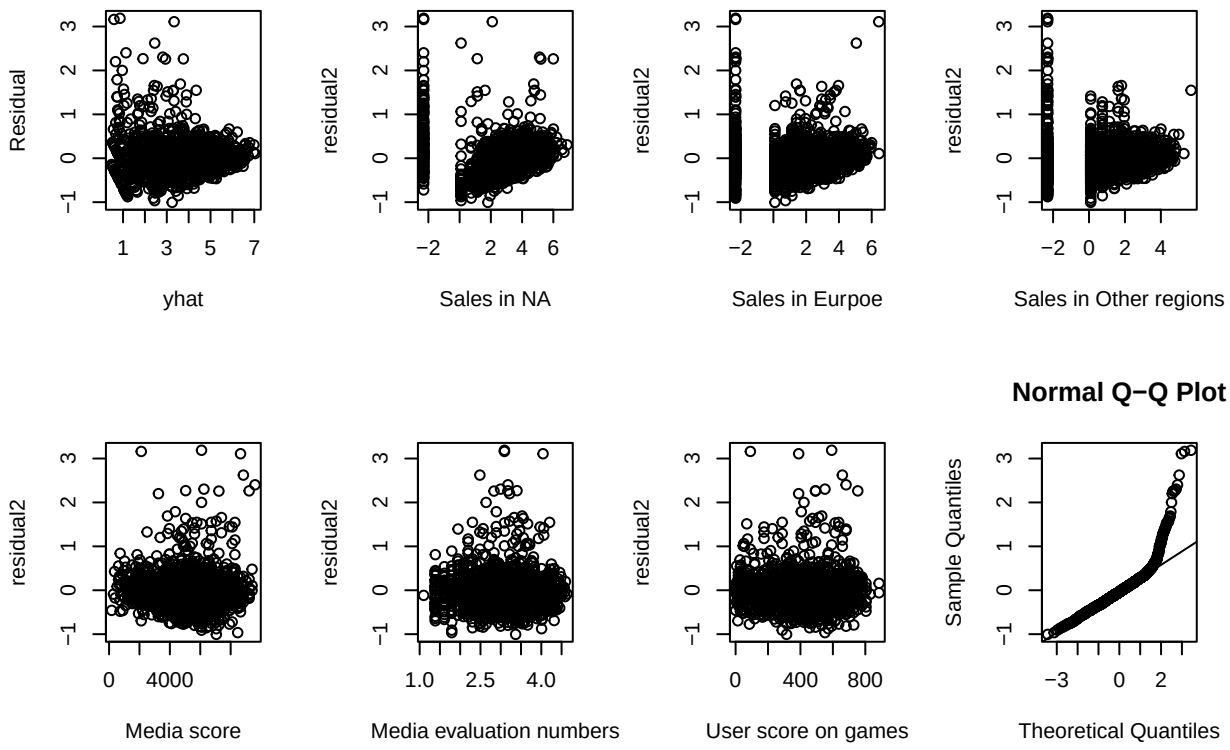


Figure 6: The residual plots and normal QQ plot in testing set



Reference

[1] Kunlin, Li. et,al.(2020, July)*Exploring the influence of online reviews and motivating factors on sales: A meta-analytic study and the moderating role of product category.*School of Information Management, Wuhan University, PR China[https://www.sciencedirect.com/science/article/pii/S0969698919304011].

[3] Mirko Ernkvist & Patrik Ström (2018) *Differentiation in digital creative industry cluster dynamics: the growth and decline of the Japanese* Geografiska Annaler: Series B, Human Geography, 100:3, 263-286, DOI: 10.1080/04353684.2017.1423506[https://www.tandfonline.com/doi/full/10.1080/04353684.2017.1423506]

[3] Neil, T. et, al.(2013, Sep) *The Impact Of Platform On Global Video Game Sales.*International Business & Economics Research Journal (IBER) [https://www.researchgate.net/publication/297754899_The_Impact_Of_Platform_On_Global_Video_Game_Sales].

[4] SID_TWR(2018) *Video Games Sales Dataset* Kaggle.com [https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset/code]